# Deep reinforcement learning and the role it played in beating Dota 2 champions.

**Enrico Dreyer**
**31210783**

**ITRI 616 Exam**

# TABLE OF CONTENTS

# LIST OF FIGURES

# ITRI 616 AI Exam

## 1. Introduction

Inventors have long desired to make machines that can think the way that humans do (Kai et al., 2013). Since programmable computers, people wondered if these computers could ever become intelligent (Goodfellow et al., 2016). In the beginning we looked at intelligent software with the intent to solve human problems, such as make medical diagnoses, routine labour automatic, have an understanding of images and speeches, as well as support scientific research (Kelleher, 2019).

We generally use artificial intelligence to solve problems that are difficult for humans, but straight forward for computers (Goodfellow et al., 2016). The real challenge for artificial intelligence is to do what a human find easy, where a computer finds it difficult, for example tasks that we as humans feel happen automatic, such as reading spoken words out loud, or recognizing faces on photos (Goodfellow et al., 2016).

In this paper, deep learning will be discussed, followed by how deep reinforcement learning applied to the development of OpenAI, an AI system that competed in a real-time Dota 2 game against professional players.

## 2. History of deep learning

According to Thomas (2020) neural networks and deep learning date back to the 1950's. It started with a british computer scientist and mathmatician Alan Turing, that predicted that humans will create a supercomputer that can act like an inteligent human. In 1986 Geoffrey Hinton ("godfater of Deep Learning") and Carnegie Mellon was some of the researchers that demonstrated that more than one neural network could be trained by using backpropagation, their study was to improve the way that an AI system can identify shapes and predict words (Thomas, 2020).

Deep learning has improved dramatically in different artificial intelligent (AI) tasks such as machine translation, speech recognition and object detection (Wang & Raj, 2017). But the very nature of the deep architecture, researchers have extended the possibility of solving a variety of modern domains that exceed the norm of basic AI tasks, for example the diagnostics of speech signals, and the use of stacked autoencoders to find clustered patterns in gene expressions (Wang & Raj, 2017).

## 3. Deep Learning

According to Kelleher (2019) deep learning is a subfield of artificial intelligence that focuses on making big neural network models that are able to make data-driven decisions accurately. Deep learning is most useful when the data is complex and where the dataset is large (Kelleher, 2019).

Deep learning takes out some of the data pre-processing that is normally involved with machine learning (Education, 2020). Algorithms that can be used, can process and ingest data that is unstructured, for example images or frames, while removing human expert dependency by automating feature extraction (Education, 2020). A good example of his is having a set of photos, where you want to categorize the photos by “dog” or “cat”. A deep learning algorithm can determine the characteristics of each animal and be able to distinguish between them (Wang & Raj, 2017).

## 4. How deep learning works

Deep learning neural networks attempt to mimic the human brain with a combination of weights, inputs, and bias (Kai et al., 2013). With the use of these elements they can work together to describe, recognize and classify objects in a certain dataset with great accuracy (Education, 2020).

Using multiple layers of interconnected nodes, deep neural networks can build upon previous layers to optimize and refine categorization or prediction (Goodfellow et al., 2016). Forward propagation is the progression of computations that is used throughout the network, where the visible layers are the input and output of the layers in a deep neural network (Education, 2020). At the input layer the deep learning model ingests data used for processing, whereas the output layer is where the final classification or prediction is made (Education, 2020).

Backpropagation is a process that uses algorithms to calculate errors in a prediction, it then adjusts the biases or weights of that function, in the effort to train that model by moving backwards through the neural network layers. Using backpropagation and forward propagation, neural networks can correct for any errors to provide a more accurate prediction over time (Education, 2020).

This is deep neural networks in the simplest terms, deep learning can become incredibly complex, as well as have more than one type of neural network, it depends on the problem that needs to be solved (Kai et al., 2013).

## 5. Deep reinforcement learning

According to Henderson et al. (2018) reinforcement learning is the study of how an agent interacts with the environment that it is put in, then learn through trial and error a policy that maximizes the expected rewards when given a certain task. This type of deep learning has shown increasing interest in the areas that include controlling continuous systems, such as robotics, playing Go, Atari and competitive video games (Henderson et al., 2018).

The reason why reinforcement learning is achieving enormous milestones is because it enables automation of end-to-end learning and feature engineering (Li, 2017). Feature engineering can be time consuming,

incomplete or over-specified, but with the use of gradient decent the process significantly reduces the reliance on domain knowledge (Li, 2017).

Deep reinforcement learning is important for the next example of deep learning, as this was the method used to train OpenAI, an AI system used to learn to play Dota 2 with the objective to compete against professional players.

## 6. Example of deep reinforcement learning

### *A. What is Dota 2?*

According to Tachintha (2020) Dota 2 is a multiplayer online battle arena (MOBA), and the abbreviation “Dota” stands for “Defence of the ancients”. The goal of the game is to defend your own “ancient”, which is a large structure in the back of your stronghold. The map layout can be found in Appendix C.

A single Dota match is played by two teams of five players, each defending their own Ancients and you win by destroying the other teams ancient (PCGamesN, 2021). Each player controls their own individual character called a “hero”, each hero has their own unique playing styles and abilities, and according to Nathan (2021) Dota 2 has 121 heroes to choose from.

During the match, players buy or collect “items” and experience points (XP) that help them in defeating the opposing team in combat (Berner et al., 2019). The player collects gold by defeating creeps (basic non-player units), destroying an enemy tower or defeat an enemy hero. Apart from learning all the abilities of each hero, there are 150 purchasable items and 58 neutral items that can be picked up by a player (Sengupta, 2020).

### *B. OpenAI*

The lifelong goal of AI is to solve real-world advanced problems. In 2016, an AI called AlphaGo defeated a world champion Go player using Monte Carlo tree search and deep reinforcement learning (Granter et al., 2017). Deep reinforcement learning does not stop at AlphaGo, but DRL models have tackled tasks like text summarization, robotic manipulation, as well as other games such as Minecraft and Starcraft (Kelvin & Schneiders, 2018).

According to Berner et al. (2019) OpenAI Five became the first AI system that could defeated the world champions in a standard ranked game of Dota 2 on April 13$^{th}$ 2019. An image of the main event can be found in Appendix G. Dota 2 represents numerical challenges for AI systems such as imperfect information, long time horizons and continuous state-action spaces.

OpenAI Five used existing reinforcement learning techniques, learning at a batch of approximately 2 million frames every 2 seconds. The OpenAI Five team developed tools and a distributed training system, that allowed the OpenAI Five to train continuously for approximately 10 months. The objective of this AI was to beat the Dota 2 world chapions, Team OG (shown in appendix F), to demonstrate that self-play reinforcement learning can perform at a superhuman level to achieve a difficult task.

Unlike Go or Chess, complex games capture the continuous nature and complexity of the real world. Dota 2 proved to be the perfect challenge as it is a multiplayer, real-time strategy game that was created by Valve in the mid 2013 (Berner et al., 2019). Dota 2 has an average player base of between 500,000 and 1,000,000, as well as having full time professionals (Berner et al., 2019). The 2019 international championship prize pool of just over $35 million which proved to be the largest prize pool in the world, at that time (Berner et al., 2019).

One of the most important parts of solving the complexity of the environment is to scale existing reinforcement learning systems to extraordinary levels that the system was not use to (Berner et al., 2019). One of the biggest challenges for the OpenAI team was the environment that kept on changing in the 10-month training cycle, as Dota 2 had weekly updates. To train the AI without having to restart the training every time the environment changed, the team developed a collection of tool that resumed the training with minimal effect to the performance, they called it surgery (Berner et al., 2019). The team performed a surgery every two weeks in the 10-month training period.

### *C. Challenges when playing Dota 2*

According to (Berner et al., 2019), for the AI system to play Dota 2 it has to overcome various challenges:

**Long-time horizons** – According to ChessGames (2021) the average chess game has 41.03 moves, whereas Dota 2 runs at 30 frames per second with an average game being 45 minutes. OpenAI Five acts out an action every fourth frame, coming to a total of approximately 20,000 moves every game (Berner et al., 2019).

**Partially observed state** – Each team can only see the portion of the map that their units, buildings, or observer wards (item that can be bought to show a small area of the map) can see, the rest of the map is hidden. This requires OpenAI Five to make inferences based on the opponent's behaviour and data that is incomplete (Berner et al., 2019).

**Observation spaces and high-dimensional action** – Variables that OpenAI needs to observe can include the ten heroes, creeps, buildings as well as game features such as trees, wards and runes. According to Berner et al. (2019) these variables can stack up to an average of 16,000 per observation, and the AI chooses one action between 8,000 and 80,000 possible actions.

The OpenAI team decided to reduce the complexity by setting limitations for its own AI, by only having it learn 17 heroes, as well as not support items that allow the AI control other units besides itself.

### *D. OpenAI training system*

#### (i) How it works

A normal human player interacts with the game using a computer monitor, mouse and keyboard. Decisions are made in real time and they reason with the long-term consequences in mind (Berner et al., 2019). Dota 2 runs at 30 frames per second, while the OpenAI Five acts on every 4$^{th}$ frame which the OpenAI team called a timestamp. For every timestamp the AI receives the information that a human player will see such as unit health, position and mana count (Berner et al., 2019). The interface of a human player can be found in Appendix D. The AI then sends an action to the game engine, giving its desired action such as move, attack or use an ability.

The OpenAI team already surpassed professional-level play by hand-scripting some game mechanics such as unique courier unit controls, which items a hero keep in its reserves, and when a hero is allowed to purchase abilities or items. Although the team believes that it can perform better without the hand-scripted logic it is still in their long-term planning (Raiman et al., 2019).

As shown in Figure 1, the OpenAI team defined a policy ($\pi$) as a function that forms part of the history of observations to a probability distribution over the AI's actions, which they parameterized as a recurrent neural network with an approximate parameter ($\theta$) count of 159 million. Primarily the neural network consists of a single-layer 4096-unit (LSTM) (Berner et al., 2019). The OpenAI team let the AI play games by repeatedly passing current observation as sampling an action as input and distributing the output at each timestamp.

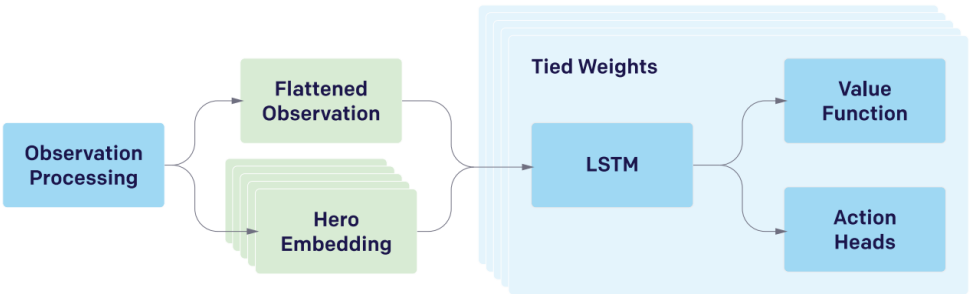


**Figure 1: Simplified OpenAI Five model Architecture (Berner et al., 2019)**

The OpenAI team (2019) explained their simplified model Architecture as a single vector that is used to process the complex multi-array observation space, when it is done, it is passed to the 4069-unit LSTM. An example graph showing the multiple arrays received at every timestamp can be found in Appendix E. The policy outputs are obtained by the LSTM state (value of each action). Each hero is controlled by a clone of the network with almost the same inputs, and each of them having its own hidden state. Different actions are taken by the network according to the observation processing's output, depending on what hero is being controlled. 84% of the total parameter count is composed by the LSTM.

(ii) Optimization of the policy

The goal of the OpenAI team is to find a policy that maximizes the probability to win against a professional human team. They used a reward function that included additional signals to indicate heroes dying, winning a lane or the collection of resources (Berner et al., 2019), the official reward system can be found in Appendix A. They also compared the resources of the other team to exploit the zero-sum multiplayer structure, as Dota 2 is a game of resources and achieving objectives, if the opposing team has a higher net worth then the chances of them winning becomes higher (Kinkade et al., 2015).

The policy used in OpenAI was trained by using Proximal Policy Optimization (PPO) (Berner et al., 2019). Generalized Advantage Estimation (GAE) was the optimization algorithm used to accelerate training and stabilization (Berner et al., 2019). The OpenAI team trained their policy using self-play experience by playing Dota 2 against itself. To achieve this, they had a central pool of optimizer GPUs that received game data and stored it asynchronously in local buffers that they called experience buffers. By using sampling minibatches from the experience buffer at random, each optimizer GPU can calculate gradients that is averaged across the pool. This means that the more GPUs they were using the faster it can process the experience buffer, thus the faster it can train (Berner et al., 2019).

"Rollout" worker machines were used to run the self-play games (Berner et al., 2019). By running games at almost half real-time speed they found that they can run more than twice as much games in parallel and increased the learning speed of the AI (Raiman et al., 2019). The OpenAI team also made use of sending smaller amounts of game data at a time, rather than sending everything at the end of a game, a graph explaining can be found in Appendix B. They played the latest policies against itself 80% of the time and an older version 20% of the time, this is to prevent strategy collapse and to obtain more robust strategies. In some cases the AI forgot how to play against different strategies and only focused on defeating its current self (Berner et al., 2019).

## 7. Conclusion

OpenAI five ended up beating team OG in a best of three with a score of 2-0. This proved that when scaled up successfully, reinforcement learning techniques can be used to achieve superhuman performances.

Raiman et al. (2019) stated that future work includes to improve on fight or flight prediction and have the AI play with other human players on the same team, as well as communication its future plans to the other human players.

In the match between OpenAI and Team OG, one of OpenAI's heroes bought back after being defeated in a situation that seemed to be a bad move according to pro players, but that hero ended up giving vision for another hero to defeat two opposing heroes. This shows that deep learning is not only there to mimic the human brain but outperform humans and dramatically improve our quality of life by helping us learn faster and understand factors that we as humans have not even thought about (Education, 2020). An AI system also had a steeper learning curve as it can understand complex patterns better than humans and being able to work longer hours without stopping (Wang & Raj, 2017).

## 8. Reference List


ListBerner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., & Hesse, C. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

ChessGames. (2021). *Statistics page*. https://www.chessgames.com/chessstats.html

Education, I. C. (2020). Deep learning. https://www.ibm.com/cloud/learn/deep-learning

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. https://books.google.co.za/books?hl=en&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=deep+learning&ots=MNO3emkANZ&sig=JK0HdYvfhLZuq9nZazgi_umoIM4#v=onepage&q=deep%20learning&f=false

Granter, S. R., Beck, A. H., & Papke Jr, D. J. (2017). AlphaGo, deep learning, and the future of the human microscopist. *Archives of pathology & laboratory medicine*, *141*(5), 619-621.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI conference on artificial intelligence,

Kai, Y., Lei, J., Yuqiang, C., & Wei, X. (2013). Deep learning: yesterday, today, and tomorrow. *Journal of computer Research and Development*, *50*(9), 1799.

Kelleher, J. D. (2019). *Deep learning*. MIT press. https://books.google.co.za/books?hl=en&lr=&id=b06qDwAAQBAJ&oi=fnd&pg=PP9&dq=deep+learning+Kelleher,+John+D&ots=_oBXSRp-_O&sig=cu8Pew7SvYofKm87MEQ6xLeBG8c#v=onepage&q=deep%20learning%20Kelleher%2C%20John%20D&f=false

Kelvin, M. J., & Schneiders, D. (2018). Learning to Play Computer Games with Deep Learning and Reinforcement Learning.

Kinkade, N., Jolla, L., & Lim, K. (2015). Dota 2 win prediction. *Univ Calif*, *1*, 1-13.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Nathan. (2021). *How Many Heroes are in Dota 2 in 2021*. https://deluxegamer.com/how-many-heroes-are-in-dota-2-2021/

PCGamesN. (2021). *How to play Dota 2 – a beginner's guide*. https://www.pcgamesn.com/dota/dota-2-beginner-s-guide-everything-you-need-know

Raiman, J., Zhang, S., & Wolski, F. (2019). Long-term planning and situational awareness in OpenAI five. *arXiv preprint arXiv:1912.06721*.
Sengupta, S. (2020). *Check out this guide if you need help understanding the items in Dota 2*. https://www.redbull.com/in-en/dota-2-items-tips-guide#:~:text=Dota%202%20has%20208%20items%20in%20total%20and,outcomes%20so%20it%20can%20be%20tough%20to%20grasp.
Tachintha, I. (2020). *Climbing the Dota 2 Difficulty Curve*. https://superjumpmagazine.com/climbing-the-dota-2-difficulty-curve-336261427586
Thomas, M. (2020). *THE HISTORY OF DEEP LEARNING: TOP MOMENTS THAT SHAPED THE TECHNOLOGY*. https://builtin.com/artificial-intelligence/deep-learning-history
Wang, H., & Raj, B. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.

## 9. Appendix A

| Name | Reward | Heroes | Description |
|---|---|---|---|
| Win | 5 | Team | |
| Hero Death | -1 | Solo | |
| Courier Death | -2 | Team | |
| XP Gained | 0.002 | Solo | |
| Gold Gained | 0.006 | Solo | For each unit of gold gained. Reward is not lost when the gold is spent or lost. |
| Gold Spent | 0.0006 | Solo | Per unit of gold spent on items without using courier. |
| Health Changed | 2 | Solo | Measured as a fraction of hero's max health.‡ |
| Mana Changed | 0.75 | Solo | Measured as a fraction of hero's max mana. |
| Killed Hero | -0.6 | Solo | For killing an enemy hero. The gold and experience reward is very high, so this reduces the total reward for killing enemies. |
| Last Hit | -0.16 | Solo | The gold and experience reward is very high, so this reduces the total reward for last hit to $\sim 0.4$. |
| Deny | 0.15 | Solo | |
| Gained Aegis | 5 | Team | |
| Ancient HP Change | 5 | Team | Measured as a fraction of ancient's max health. |
| Megas Unlocked | 4 | Team | |
| T1 Tower* | 2.25 | Team | |
| T2 Tower* | 3 | Team | |
| T3 Tower* | 4.5 | Team | |
| T4 Tower* | 2.25 | Team | |
| Shrine* | 2.25 | Team | |
| Barracks* | 6 | Team | |
| Lane Assign† | -0.15 | Solo | Per second in wrong lane. |

**Figure 2: official reward system for OpenAI Five (Berner et al., 2019)**

All five heroes (agents) have one goal and that is to win the game. Each agent gets either a reward or penalty based on what humans that play the game professionally generally agree on what is bad or good.

## 10. Appendix B

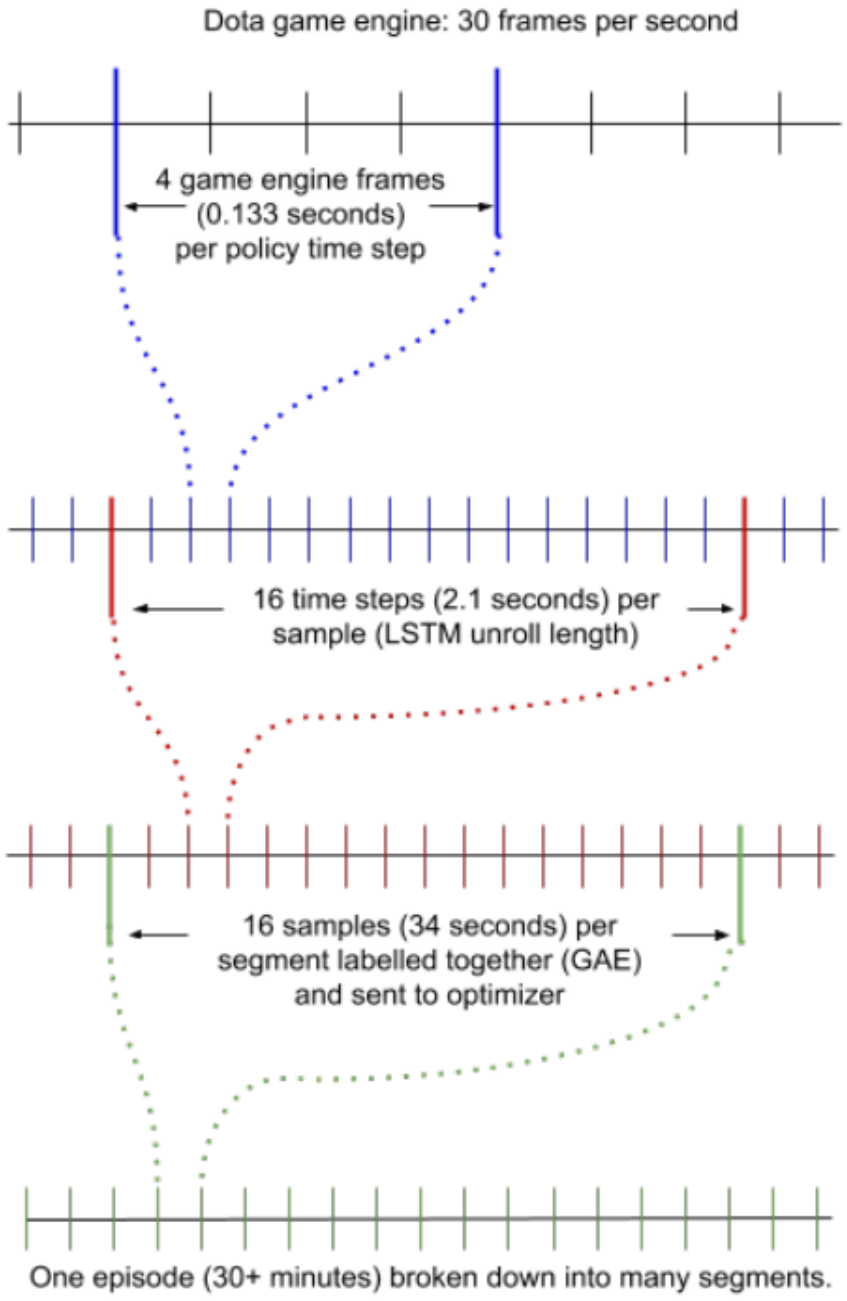


**Figure 3: Timescales and Staleness (Berner et al., 2019)**

This figure represents the way in which the optimizers receive data from the game. Rather than sending data after a whole game is played, the rollout machines send data in shorter terms.

## 11. Appendix C

**Figure 4: Dota 2 Map (own example)**

The map consists of 3 lanes:

- Top lane called “off lane” and is played by 2 heroes (one carry and one support).
- Middle lane and is played by 1 hero (core player)
- Bottom lane called “safe lane” and is played by 2 heroes (one carry and one support).

## 12. Appendix D


**Figure 5: Human Obervation Space (Berner et al., 2019)**

As shown in Figure 5, this is the normal user interface, but OpenAI uses a more semantic observation space, because its goal is to study strategic gameplay and planning rather than visual processing and the AI will waste resources on rendering the frames of each game.

## 13. Appendix E

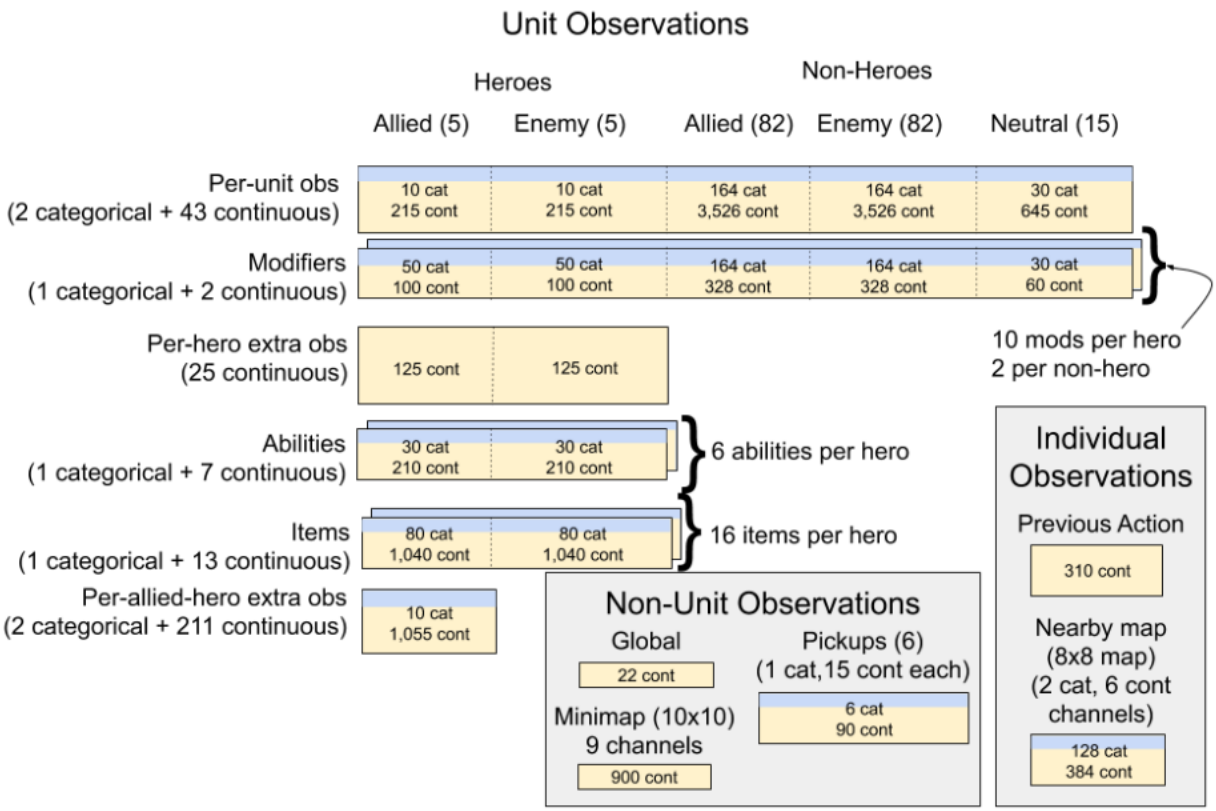


**Figure 6: Observation Space Overview**

This is a visual of each array at every timestep observed by OpenAI Five.

## 14. Appendix F

**Figure 7: Team OG Players**

Team OG is a professional esports team based in Europe, that won both The International 2018 and 2019. From left to right is Ceb (Sébastien Debs), Ana (Anathan Pham), n0tail (Johan Sundstein), Topson (Topias Taavitsainen) and JerAx (Jesse Vainikka)

## 15. Appendix G


**Figure 8: OpenAI vs Team OG**

On the 13th of April 2019 Team OG lost against OpenAi in a best of 3. In the second map Team OG lost in under 20 minutes.